Theory-Based Induction

Charles Kemp (ckemp@mit.edu)
Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139-4307 USA

Abstract

We show how an abstract domain theory can be incorporated into a rational statistical model of induction. In particular, we describe a Bayesian model of category-based induction, and generate the prior distribution for our model using a formal theory of the distribution of biological properties across classes of biological kinds. Our theory-based model is both more principled than previous approaches and better able to account for human ratings of argument strength.

Philosophers since Hume have struggled with the logical problem of induction, but children solve an even more difficult task — the practical problem of induction. Children somehow manage to learn concepts, categories, and word meanings, and all on the basis of a set of examples that seems hopelessly inadequate. The practical problem of induction does not disappear with adolescence: adults face it every day whenever they make any attempt to predict an uncertain outcome. Inductive inference is a fundamental part of everyday life, and a fundamental phenomenon in need of a psychological explanation.

Two important questions can be asked about inductive generalization: what resources does a person bring to an inductive task, and how are these resources combined to generate a response to the demands of the task? In other words, what is the process of induction, and what is the prior knowledge required by this process? Psychologists have considered both of these questions in depth, but previous computational models of induction have tended to emphasize process to the exclusion of prior knowledge. This paper attempts to redress this imbalance by showing how prior knowledge can be included in a computational model founded on rational statistical inference.

The importance of prior knowledge has been attested by psychologists and machine learning theorists alike. Murphy and Medin (1985) have suggested that the acquisition of new concepts is guided by "theories" — networks of explanatory connections between existing concepts. Machine learning theorists have built formal models of learning, and argued that generalization within these models is not possible unless a learner begins with some sort of inductive bias (Mitchell, 1997). The challenge that inspires our work is to develop a model with an inductive bias that is well motivated by a theory of the domain under consideration.

Many previous models have taken similarity judgments as their representation of prior knowledge (Nosofsky, 1986; Osherson et al., 1990). This approach has been dominant within the tradition of category-based induction, and Osherson et al.'s (1990) similarity-coverage model will be one standard against which our new model will be compared. Using similarity data to represent prior knowledge is a reasonable first attempt, but similarity judgments are less than ideal as a starting point for a model of inductive inference. As Goodman (1972) has pointed out, similarity is a vague and elusive notion. It is meaningless to say that two objects are similar unless a respect for similarity has been specified. Any model based on similarity alone is therefore a model without a secure foundation.

Instead of relying on similarity, the model developed in this paper is founded on a simple theory of a particular domain of reasoning: kinds of animals and their properties. The theory consists of two components: the 'taxonomic principle,' which holds that the set of animals naturally forms a tree-structured taxonomy, and the 'distribution principle,' which specifies how properties are probabilistically distributed over the taxonomy. These two principles are used to generate a prior distribution for a Bayesian model of category-based induction.

Our approach is located at the most abstract of Marr's three levels: the level of computational theory (Marr, 1982). Our goal is not to describe the process by which people make inductive inferences, but rather to explain why people reason the way that they do and how they can reliably come to true beliefs about the world from very limited data. Intriguingly, both the taxonomic principle and the distributional principle resemble analogous principles in evolutionary biology and genetics. People's remarkable ability to make successful inductive leaps may thus be explained as the result of rational inference mechanisms operating under the guidance of a domain theory that reflects the true structure of the

environment.

We begin by introducing previous approaches to the problem of category-based induction. We then set out a 'theory of biological properties' that can generate the prior distribution for a Bayesian model of induction. Next we turn to experimental data, and show that our new model performs better than previous approaches across a collection of four data sets. We conclude by discussing ways in which Bayesian and traditional similarity-based approaches might be complementary, and directions for future work.

Category-Based Induction

The tasks to be considered were introduced by Osherson et al. (1990). In the first task (the specific inference task), subjects are asked to rate the strength of arguments of the following form:

Horses can get disease X
Cows can get disease X

Dolphins can get disease X

The premises state that one or more specific mammals can catch a certain disease, and the conclusion (to be evaluated) states that another specific species (here dolphins) can also catch the disease.

In the second task (the general inference task), subjects are asked to consider a generalization from specific premises to a property of all mammals. For instance:

Seals can get disease X
Dolphins can get disease X

All mammals can get disease X

Previous Models

Similarity-based models. Osherson's similarity-coverage model expresses the strength of an argument as a linear combination of two components: a term representing the similarity between the premises and the conclusion, and a term representing the extent to which the premises cover the lowest level taxonomic category including both premises and conclusion.

Formalizing these ideas, the strength of the argument from a set of premises X to a conclusion category Y is:

$$\alpha\ \texttt{setsim}(X, Y) + (1 - \alpha)\ \texttt{setsim}(X, [X; Y])$$

where α is a free parameter, $\texttt{setsim}(\cdot)$ is a setwise similarity metric, and $[X; Y]$ is the lowest level taxonomic category including X and Y.

Several setwise similarity metrics might be tried. Osherson et al. propose $\texttt{maxsim}(\cdot)$ but also consider $\texttt{sumsim}(\cdot)$:

$$\texttt{maxsim}(X, Y) = \sum_j \max_i \texttt{sim}(X_i, Y_j)$$

$$\texttt{sumsim}(X, Y) = \sum_j \sum_i \texttt{sim}(X_i, Y_j).$$

Both are defined in terms of $\texttt{sim}(\cdot)$, the standard pairwise similarity metric, assumed as a primitive.

The similarity-based approach can offer no principled reason for preferring one of these metrics. Osherson et al. suggest that $\texttt{maxsim}(\cdot)$ conforms better to their intuitions, yet $\texttt{sumsim}(\cdot)$ is more standard in models of inductive learning, categorization, and memory. Later we show that $\texttt{maxsim}(\cdot)$ fits the experimental data much better than $\texttt{sumsim}(\cdot)$ and offer a possible explanation for its success.

Bayesian Models. Heit (1998) and Sanjana and Tenenbaum (2003) considered Bayesian approaches to category-based induction. Assume that we are working within a finite domain. For the tasks modeled here, the domain will be a set of ten mammal kinds. We are interested in a concept, C, that picks out some subset of these objects. In the examples above, C is the concept "mammals that can get disease X." Let H be our hypothesis space: the set of all possible concepts over our domain. With 10 animal types, there are 2^{10} distinct subsets of animals, or logically possible concepts. To each hypothesis h in H we assign a prior probability $p(h)$, where $p(h)$ is the probability that h is the concept of interest.

Osherson's first task may now be formalized as follows. We observe X, a set of n examples of the concept C, and want to compute $p(y \in C|X)$, the probability that another object, y, is also a member of C. Summing over all hypotheses in H, we have:

$$p(y \in C|X) = \sum_{h \in H} p(y \in C, h|X) \quad (1)$$

$$= \sum_{h \in H} p(y \in C|h, X)p(h|X). \quad (2)$$

Now $p(y \in C|h, X)$ equals one if $y \in h$ and zero otherwise (independent of X). Thus:

$$p(y \in C|X) = \sum_{h \in H: y \in h} p(h|X) \quad (3)$$

$$= \sum_{h \in H: y \in h} \frac{p(X|h)p(h)}{p(X)} \quad (4)$$

where the last step follows from Bayes' rule.

The numerator in Equation 4 depends on $p(X|h)$, the likelihood of X given h, as well as on the prior $p(h)$. Assuming the n examples in X are sampled independently at random from h yields:

$$p(X|h) = \begin{cases} \frac{1}{|h|^n}, & \text{if all } n \text{ examples in } X \text{ belong to } h \\ 0, & \text{otherwise} \end{cases}$$

where $|h|$ denotes the size of h (the number of instances in its extension). The denominator in Equation 4 can be computed by summing over all hypotheses: $p(X) = \sum_{h \in H} p(X|h)p(h)$.

Osherson's general inference task is formulated similarly. The probability that all of the members of category Y belong to C is:

$$p(Y \subset C|X) = \sum_{h \in H} p(Y \subset C|h, X)p(h|X) \quad (5)$$

$$= \sum_{h \in H: Y \subset h} p(h|X). \quad (6)$$

The only piece missing from the Bayesian framework is a specification of how the prior probabilities $p(h)$ are calculated. Heit (1998) does not address this question, and Sanjana and Tenenbaum (2003) use a prior distribution that is not deeply motivated by a theory of the domain. We now describe a principled method for generating the prior distribution.

A Theory-Based Model

The prior distribution for our Bayesian model is motivated by two principles: the 'taxonomic principle' and the 'distribution principle.' Together these principles form a theory of the distribution of biological properties.

The taxonomic principle holds that animals naturally fall into a tree-structured taxonomy – a collection of hierarchical groups. This belief may well be universal. A substantial body of work has documented that cultures all over the world organize living kinds into 'folk taxonomies' (Atran, 1995). It is also scientifically sound, as the theory of evolution implies that living kinds should conform to a tree structure.

Our first step towards generating a prior distribution is therefore to build a folk taxonomy for the ten mammals in our domain. Osherson collected similarity ratings between all pairs of animals in the domain, and we use these ratings to define a distance measure d, where $d(x, y) = 1 - \texttt{sim}(x, y)$. We then perform average-link clustering, which first assigns each animal to its own cluster, then proceeds by repeatedly merging the two closest clusters. The tree produced is shown in Figure 1.

Although our distance measure is defined in terms of similarity, our approach does not depend critically on similarity as a primitive. We could use other measures of distance: for example, one could represent each animal using a set of behavioral and morphological features (e.g., 'lives in water,' 'has a tail'), and set the distance between two animals to be the distance between their feature vectors. We could also use more structured domain knowledge that might obviate the need for any bottom-up clustering. Building the taxonomy without reference to similarity is our preferred approach, but using Osherson's similarity data yields one important payoff for the present study: it allows the performance of our model to be directly compared with the performance of the similarity-based models.

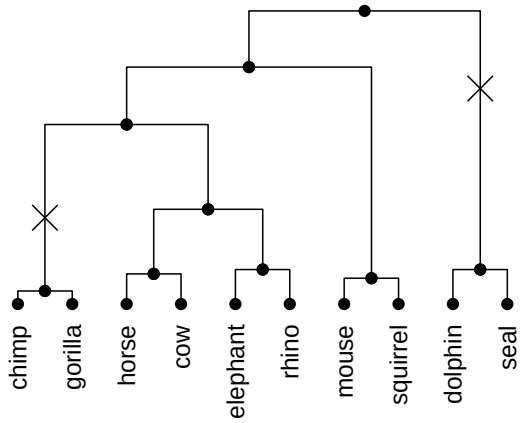

Figure 1: A taxonomy built from similarity data

A simple prior distribution can be generated using the taxonomy alone. There are 19 nodes in the tree (one for each animal type, together with nine internal nodes), and each specifies a concept that includes the animals falling beneath it on the tree. A straightforward way to set the prior is to assign a uniform probability to each of these 19 concepts, and zero probability to all other possible concepts. We call the model that uses this prior the 'Taxonomic Bayesian model.'

The Taxonomic model is appealingly simple and corresponds roughly to some informal accounts of taxonomically driven induction (e.g., Atran, 1995). But to see that it is not adequate, compare the argument "seals and squirrels catch disease X, so horses are also susceptible" with "seals and cows catch disease X, so horses are also susceptible." The second is stronger than the first, yet the Taxonomic model assigns them both the same rating, since each set of premises is compatible with only one hypothesis, the set of all mammals.

The distribution principle, the second part of our theory, states that concept membership (or feature possession) depends on a process of random mutation operating over the taxonomy. This principle acknowledges that convergent evolution can occur: that two animals may share a property even if it occurs in none of their common ancestors. Some additional notation is needed to make this principle precise.

Imagine that membership in C, the concept to be learned, depends on a single feature F that could have evolved at any point in the tree and may have evolved independently along different branches of the tree. Once F has arisen along a branch, all nodes falling below that branch are members of C. For example, if F appears at the points marked with crosses in Figure 1, then C will include chimps, gorillas, dolphins and seals.

We model the evolution of F using a Poisson arrival process with a free parameter, λ that will be

called the mutation rate. The Poisson process assumes that mutations arrive randomly, potentially occurring at any point along any tree branch, but are constrained to respect an overall rate of λ. The probability that the feature develops along a branch b of length $|b|$ is one minus the probability that the mutation arrives zero times along that branch:

$$p(F \text{ develops along } b) \quad = \quad 1 - e^{-\lambda |b|}. \qquad (7)$$

A branch is 'marked' if F develops along that branch.

To obtain a single estimate of the extension of C, we consider all branches in the tree, label each as marked or unmarked according to Equation 7, then collect all external nodes that fall beneath a marked branch. Repeating this many times generates a prior distribution over all possible concepts, where the probability assigned to a concept is the proportion of times it was chosen as our estimate of C.

This prior distribution may also be computed analytically. First consider all single branches in the tree. For each branch, calculate the probability that F arises along that branch and nowhere else in the tree, and add this probability to the prior for the corresponding concept. Continue by considering all pairs of branches, all triples, and so on.

Our model of the evolution of F captures two important intuitions. First, F is more likely to develop along the longer branches of the tree. Second, since F develops independently along different branches and the probability of arising on any branch is small, the model favors simpler hypotheses — hypotheses consisting of fewer rather than more clusters.

An alternative prior over all possible concepts was considered by Sanjana and Tenenbaum (2003). They also compute $p(h)$ by taking disjunctions of the 19 hypotheses represented by the folk taxonomy. The 19 original hypotheses are assigned the highest prior probability, disjunctions of two of these are assigned a somewhat smaller probability, and disjunctions of three hypotheses are assigned a still smaller probability. This approach represents a general strategy for expanding any hypothesis space, and can be applied to hypothesis spaces that have nothing at all to do with taxonomies. Generality, however, is bought at a price: unlike our new 'Evolutionary' model, the 'Disjunctive Bayes' model of Sanjana and Tenenbaum is not deeply motivated by the structure of the domain. A symptom of this lack of principled motivation is that Disjunctive Bayes does not take the branch lengths of the taxonomic tree into account, and thus fails to predict important effects of typicality that we describe below.

Performance of the Models

Each model presented in the previous section can be used to rank a set of arguments in order of increasing strength. Figure 2 shows how well these ranks match the ranks assigned by humans.

The first three columns show the performance of the models on three data sets published in previous studies. The Osherson general set contains 45 three-premise general arguments. The Osherson specific set contains 36 two-premise arguments, and the Sanjana set contains 28 specific arguments with a variable number of premises.

All of the models (except Taxonomic Bayes) include a single free parameter, and each correlation in Figure 2 is shown for the setting of the parameter that best fits the human data. As expected, Taxonomic Bayes performs poorly, but the other two Bayesian models both outperform the similarity-based models over the first three datasets. Both of these Bayesian models show robust performance across a range of parameter settings, and both admit a single setting that achieves correlations exceeding 0.9 on the first three data sets.

A New Experiment. A limitation of the Disjunctive Bayes model is that it does not capture at least one phenomenon documented by Osherson et al. (1990). General arguments tend to increase in strength as the premises become more typical of the conclusion category. For example, since horses are more typical mammals than seals,

$$\frac{\text{Horses can get disease X}}{\text{All mammals can get disease X}}$$

is a stronger argument than

$$\frac{\text{Seals can get disease X}}{\text{All mammals can get disease X}}$$

Although the Evolutionary model was not built with premise typicality in mind, we collected new data which show that it captures this effect more successfully than the Disjunctive Bayes model. Ten single-premise general arguments (one for each species in our domain) were printed on a set of cards, and 25 undergraduates sorted these cards in order of increasing argument strength. The average rank of each argument was calculated and compared with the ranks assigned by the models. Owing to the limited number of arguments, correlations are much lower than for the previous three data sets. Figure 2 nonetheless shows that the Evolutionary Bayes and similarity models partially capture the premise typicality effect, but the other Bayesian models do not.

Evolutionary Bayes and Maxsim

Sumsim performs dramatically worse than Maxsim, to the point of being anticorrelated with people's judgments on the Osherson general stimuli. This confirms the intuition that `maxsim`$(\cdot)$ is the better metric for category-based induction, but the superiority of Maxsim still awaits a principled explanation.

Heit (1998) and Sanjana and Tenenbaum (2003) have suggested that Bayesian analyses might explain

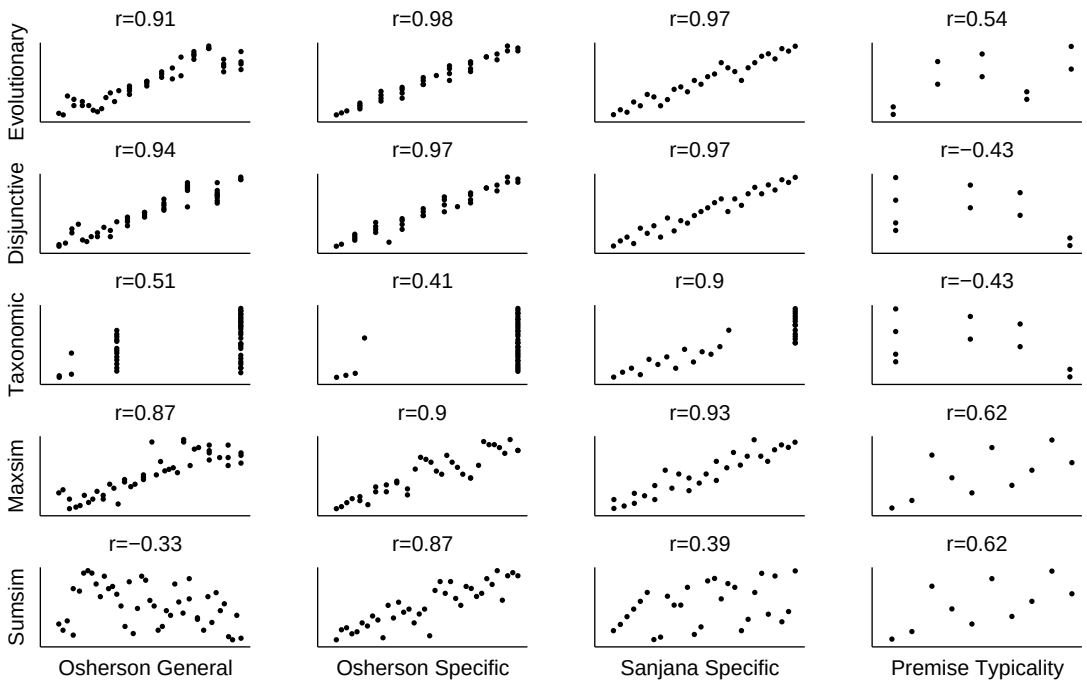

Figure 2: Model predictions (x axis) vs human judgments (y axis). Each row shows the performance of a model, and each column shows the performance over a data set. Every model (except Taxonomic Bayes) includes a single free parameter, and the plots shown are for the best setting of that parameter.

the success of other approaches to category-based induction. Noticing that Maxsim and Evolutionary Bayes perform similarly on all four data sets, we conjectured that Maxsim might effectively be an approximation to the more principled but more complex Evolutionary Bayes model. Such a correspondence would support a rational justification for why human inference might follow the Maxsim rule in this domain.

To further explore the relationship between these two models, we ran a simulation using a set of 100 randomly generated taxonomies. Each taxonomy was generated by starting with a set of 10 nodes, and merging pairs of nodes at random until only one remained. The branch lengths were generated by choosing 9 random numbers between 0 and 1, and setting the height of the node created by the kth merge to the kth smallest of these numbers. For each taxonomy, we calculated the correlations between the predictions of each pair of models on an analog of the Osherson general task. To calculate the predictions of the similarity models, the similarity of two objects was defined to be one minus the length of the path joining the objects in the tree. This makes sense under the assumption that the tree approximates the structure used to generate the similarity judgments.

Figure 3 shows the outcome of this simulation. The pair of models that matched most closely on these general arguments was Maxsim and Evolutionary Bayes, even though Evolutionary Bayes is superficially much more similar to Disjunctive Bayes than Maxsim. This result supports the idea that Maxsim and Evolutionary Bayes may specify a very similar mapping between their input (the similarity matrix) and their output (ratings of argument strength).

Marr (1982) proposed three broad levels at which a psychological account may be situated. The Bayesian approach is best suited for the formulation of models at the most abstract level of "computational theory." In contrast, Maxsim falls most naturally into Marr's second level as an algorithm that might implement the computational theory in a psychologically plausible way. The similar performance of these two models supports the idea that they are complementary. Evolutionary Bayes helps to show why Maxsim may be a reasonable model of inductive generalization, and Maxsim provides an existence proof that the computations required by the Bayesian model can be approximated by simple heuristics.

Discussion

The prior distribution used by the Evolutionary model follows directly from the theory consisting of the taxonomic and distribution principles. It is striking that a model inspired by ideas about random mutation and convergent evolution can predict people's intuitive judgments so well, but this should not be surprising if we believe that the success of

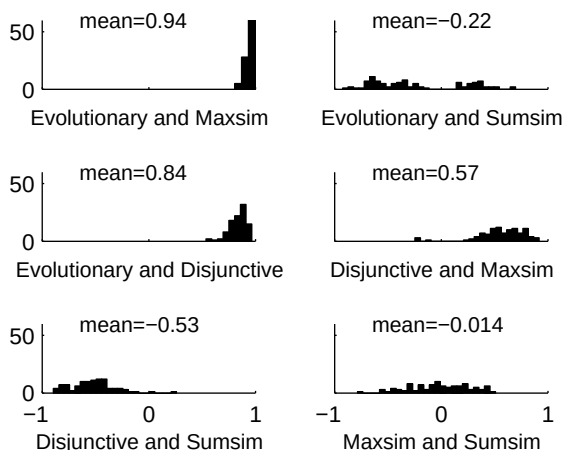

Figure 3: Frequencies (y-axis) against correlations (x-axis). Each histogram shows the distribution of the correlations achieved by a pair of models across 100 random trees.

human cognition rests on our ability to abstract the deep enduring structure of our environment. It is an open question whether the biological principles guiding our model are explicitly represented in people's minds, or only implicitly present in the inference procedures they use. Studies with experts in biological taxonomy, creationists, or other groups of individuals who at a conscious level hold different theories of biology may yield some insight into this question.

Many aspects of intuitive (and scientific) theories of biology have not been included in our model so far, but will need to be incorporated as we enlarge its explanatory scope. We have assumed that all features are independent and are equally likely to arise anywhere in the taxonomy. In contrast, both scientific and intuitive biological theories posit that features are not independent but causally related, and that causally deeper features tend to arise higher up in the taxonomy (Atran, 1995). Also, many features are not taxonomically organized, but depend on an animal's ecological niche — its behaviors and its interactions with other species. We are currently studying how to incorporate these principles into our approach by adopting more sophisticated mutation-and-selection models and alternative structures for hypotheses organized around causal networks or spaces of behavioral traits.

A theory-based approach should not be criticized if it fails to generalize beyond the domain for which it was designed. The main reason for wanting to model a theory is that domain-specific knowledge is likely to be important. Still, we are optimistic that the theory used to build our model will be useful beyond the domain of animals. It should apply to all living kinds, and more generally to any set of objects that can be represented by a developmental tree. Artifacts are one example of a non-biological domain that may meet this condition. Consider, say, the set of all electronic devices. Any two devices that share a QWERTY keyboard are similar partly because both grew out of a single previous technology.

Our behavioral experiment explored typicality effects, which have often been interpreted as evidence against taxonomically structured representations with all-or-none concepts and in favor of more graded, similarity-based representations. We showed that typicality effects in inductive reasoning are in fact compatible with a taxonomy of all-or-none concepts, under the appropriate inference engine and prior probability distribution. Typicality may arise not from the intrinsic format of the knowledge representation, but from the inferential processes operating over that representation.

More generally, by allowing a natural combination of structured domain knowledge with probabilistic inference, our Bayesian framework offers an alternative to the traditional debates of "structure versus statistics" that have polarized much of cognitive science. From a rational functional standpoint, "structure" and "statistics" are not competitors. Inductive inference should be most successful when it brings the most powerful inferential machinery together with the most accurate domain knowledge. Here we have worked out one version of this rational approach to integrating a domain theory with statistical inference, and shown that it provides a satisfying account of people's inductive judgments about animal properties.

Acknowledgments Thanks to Neville Sanjana for contributing code and unpublished results, Liz Baraff for collecting data, and Tom Griffiths, Tevye Krynski and an anonymous reviewer for valuable suggestions.

References

Atran, S. (1995). Classifying nature across cultures. In Smith, E. E. and Osherson, D. N., editors, *An Invitation to Cognitive Science*, volume 3. MIT Press.

Goodman, N. (1972). Seven strictures on similarity. In *Problems and Projects*. Bobbs-Merril Co., Indiana.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In Oaksford, M. and Chater, N., editors, *Rational Models of Cognition*, pages 248–274. Oxford University Press.

Marr, D. (1982). *Vision*. W. H. Freeman.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57.

Osherson, D., Smith, E. E., Wilkie, O., López, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, 97(2):185–200.

Sanjana, N. E. and Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Processing Systems 15*. MIT Press.